1. TopN

现在有这样一份数据：exercise_topn.txt

```
1,huangxiaoming,45,a-c-d-f
2,huangzitao,36,b-c-d-e
3,huanglei,41,c-d-e
4,liushishi,22,a-d-e
5,liudehua,39,e-f-d
6,liuyifei,35,a-d-e
```

字段的意义：

```
id,name,age,favors
id,姓名,年龄,爱好
```

其中需要注意的是：每一条记录中的爱好有多个值，以"-"分隔

需求：

```
求出每种爱好中，年龄最大的两个人（爱好，年龄，姓名）注意思考一个问题：如果某个爱好中的第二大年龄
有多个相同的怎么办？
```

```
a   huangxiaoming   45
a   liuyifei        35
b   huangzitao   36
c   huangixaoming   45
c   huanglei        41
```

```
思路总结：
1、explode() + lateral view
2、求TopN + row_number()
```

解题：

第一步：建表导入数据相关准备：

```
create database if not exists exercise_db;
use exercise_db;
drop table if exists exercise_topn;
create table exercise_topn(id int, name string, age int, favors string) row
format delimited fields terminated by ",";
load data local inpath "/home/bigdata/exercise_topn.txt" into table
exercise_topn;
select * from exercise_topn;
desc exercise_topn;
```

第二步：思路分析，如果每个人的爱好，都可以一行一个来表示，那么就很容易求了。
需要把这种数据：

```
6,liuyifei,35,a-d-e
```

变成：

```
6,liuyifei,35,a
6,liuyifei,35,d
6,liuyifei,35,e
```

Hive内置函数中，有一个explode函数：

```
explode(a) -
separates the elements of array a into multiple rows,  把一个数组编程多行一列
    or
separates the elements of a map into multiple rows and columns  把一个字典变成多行
两列
```

SQL测试实现的结果：

```
select explode(split("a-d-e", "-"));        √√√√√√√√
select explode(split(favors, "-")) from exercise_topn;      √√√√√√√√
select id, name, age, explode(split(favors, "-")) from exercise_topn;
xxxxxxx
```

为什么上面的第3个SQL语句不能执行？必须要借助于虚拟视图技术：

```
leteral view
```

改写：

```
select a.id as id, a.name as name, a.age as age, favor_view.favor
from exercise_topn a
LATERAL VIEW explode(split(a.favors, "-")) favor_view as favor;
```

得到结果：

```
+-----+----------------+------+-------------------+
| id  |      name      | age  | favor_view.favor  |
+-----+----------------+------+-------------------+
| 1   | huangxiaoming  | 45   | a                 |
| 1   | huangxiaoming  | 45   | c                 |
| 1   | huangxiaoming  | 45   | d                 |
| 1   | huangxiaoming  | 45   | f                 |
| 2   | huangzitao     | 36   | b                 |
| 2   | huangzitao     | 36   | c                 |
| 2   | huangzitao     | 36   | d                 |
| 2   | huangzitao     | 36   | e                 |
| 3   | huanglei       | 41   | c                 |
| 3   | huanglei       | 41   | d                 |
| 3   | huanglei       | 41   | e                 |
| 4   | liushishi      | 22   | a                 |
| 4   | liushishi      | 22   | d                 |
| 4   | liushishi      | 22   | e                 |
| 5   | liudehua       | 39   | e                 |
| 5    | liudehua        | 39    | f                  |
| 5    | liudehua        | 39    | d                  |
| 6    | liuyifei        | 35    | a                  |
| 6    | liuyifei        | 35    | d                  |
| 6    | liuyifei        | 35    | e                  |
+------+----------------+-------+-------------------+
```

第三步：使用普通的分组聚合技巧就可以求得每种爱好中年龄最大的一个人

```
select aa.favor, max(aa.age) as maxage
from
(
select a.id as id, a.name as name, a.age as age,  favor_view.favor
from exercise_topn a
LATERAL VIEW explode(split(a.favors, "-")) favor_view as favor
) aa
group by aa.favor;
```

结果：

```
+-----------+---------+
| aa.favor  | maxage  |
+-----------+---------+
| c         | 45      |
| f         | 45      |
| a         | 45      |
| d         | 45      |
| b         | 36      |
| e         | 41      |
+-----------+---------+
```

```
a,huangxiaoming,45
a,huangbo,43
a,huanglei,43
a,huangzitao,40
b,liushishi,22
b,liuyifei,21
b,liujialing,20
```

```
select * from (
    select aa.id, aa.name, aa.age, aa.favor,
    row_number() over (partition by aa.id order by aa.age desc) as rank
    from
    (select a.id as id, a.name as name, a.age as age, favor_view.favor as favor
     from exercise_topn a
     LATERAL VIEW explode(split(a.favors, "-")) favor_view as favor
    ) aa
) bb where bb.rank <= 2;
```

```
a,huangxiaoming,45，1
a,huangbo,43，2
a,huanglei,41，3
a,huangzitao,40，4
b,liushishi,22，1
b,liuyifei,21，2
b,liujialing,20，3
```

```
select * from table where rank <=2 ;
```

第四步：使用求解TopN的技巧就可以求得每种爱好中年龄最大的两个人

先添加序号：

```
select aa.id, aa.name, aa.age, aa.favor,
row_number() over (distribute by aa.favor sort by aa.age desc) as index
from
(
select a.id as id, a.name as name, a.age as age, favor_view.favor
from exercise_topn a
LATERAL VIEW explode(split(a.favors, "-")) favor_view as favor
) aa ;
```

得到结果数据：

```
+--------+----------------+---------+-----------+--------+
| aa.id  |    aa.name     | aa.age  | aa.favor  | index  |
+--------+----------------+---------+-----------+--------+
| 1      | huangxiaoming  | 45      | c         | 1      |
| 3      | huanglei       | 41      | c         | 2      |
| 2      | huangzitao     | 36      | c         | 3      |
| 1      | huangxiaoming  | 45      | f         | 1      |
| 5      | liudehua       | 39      | f         | 2      |
| 1      | huangxiaoming  | 45      | a         | 1      |
| 6      | liuyifei       | 35      | a         | 2      |
| 4      | liushishi      | 22      | a         | 3      |
| 1      | huangxiaoming  | 45      | d         | 1      |
| 3      | huanglei       | 41      | d         | 2      |
| 5      | liudehua       | 39      | d         | 3      |
| 2      | huangzitao     | 36      | d         | 4      |
| 6      | liuyifei       | 35      | d         | 5      |
| 4      | liushishi      | 22      | d         | 6      |
| 2      | huangzitao     | 36      | b         | 1      |
| 3      | huanglei       | 41      | e         | 1      |
| 5      | liudehua       | 39      | e         | 2      |
| 2      | huangzitao     | 36      | e         | 3      |
| 6      | liuyifei       | 35      | e         | 4      |
| 4      | liushishi      | 22      | e         | 5      |
+--------+----------------+---------+-----------+--------+
```

然后最终SQL：

```
select c.id, c.name, c.age, c.favor
from
(
select b.id, b.name, b.age, b.favor,
row_number() over (partition by b.favor order by b.age desc) as rank
from
(
select a.id as id, a.name as name, a.age as age,  favor_view.favor
from exercise_topn a
LATERAL VIEW explode(split(a.favors, "-")) favor_view as favor
) b
) c
where c.rank <= 2;
```

使用 with 语法改写：

```
with
b as (select a.id as id, a.name as name, a.age as age,  favor_view.favor
from exercise_topn a LATERAL VIEW explode(split(a.favors, "-")) favor_view as
favor),
c as (select b.id, b.name, b.age, b.favor,
row_number() over (partition by b.favor order by b.age desc) as rank
from b)
select c.id, c.name, c.age, c.favor from c where c.rank <= 2;
```

最终结果数据：

```
+-------+----------------+--------+----------+
| c.id  |     c.name     | c.age  | c.favor  |
+-------+----------------+--------+----------+
| 1     | huangxiaoming  | 45     | c        |
| 3     | huanglei       | 41     | c        |
| 1     | huangxiaoming  | 45     | f        |
| 5     | liudehua       | 39     | f        |
| 1     | huangxiaoming  | 45     | a        |
| 6     | liuyifei       | 35     | a        |
| 1     | huangxiaoming  | 45     | d        |
| 3     | huanglei       | 41     | d        |
| 2     | huangzitao     | 36     | b        |
| 3     | huanglei       | 41     | e        |
| 5     | liudehua       | 39     | e        |
+-------+----------------+--------+----------+
```